Final Project Proposal

Group 12

Noah Olsen, Jia Chen, Jianing Wang

Problem

As the use of credit cards become more convenient and offer better rewards than debit cards, credit cards became a more common method of payment in everyday use. But an impact of this method is, the more people use credit cards, the more problems come out. One of the problems is defaulting payments. If there was a method to predict the possibility of a consumer defaulting on payments, using computationally guided discovery, then this would save the company a tremendous amount of loss through proactive reporting.

This study aims to predict the possibility of default payments of credit card clients. To solve this problem, we plan to use three different approaches (Support Vector Machine, Random Forest and Naïve Bayes Model) will be used to forecast the default and the predictive accuracy of the three methods will be compared.

Data and Network

Our dataset is looking at defaults on credit card payments in Taiwan. Taken from the UCI Machine Learning Repository at this link. The dataset has 30,000 observations and includes 23 explanatory variables along with a binary target variable in the last column on

	count	mean	std	min	25%	50%	75%	max
ID	30000.0	15000.500000	8660.398374	1.0	7500.75	15000.5	22500.25	30000.0
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
SEX	30000.0	1.603733	0.489129	1.0	1.00	2.0	2.00	2.0
EDUCATION	30000.0	1.853133	0.790349	0.0	1.00	2.0	2.00	6.0
MARRIAGE	30000.0	1.551867	0.521970	0.0	1.00	2.0	2.00	3.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_0	30000.0	−0.016700	1.123802	−2.0	−1.00	0.0	0.00	8.0
PAY_2	30000.0	−0.133767	1.197186	−2.0	−1.00	0.0	0.00	8.0
PAY_3	30000.0	−0.166200	1.196868	−2.0	−1.00	0.0	0.00	8.0
PAY_4	30000.0	−0.220667	1.169139	−2.0	−1.00	0.0	0.00	8.0
PAY_5	30000.0	−0.266200	1.133187	−2.0	−1.00	0.0	0.00	8.0
PAY_6	30000.0	−0.291100	1.149988	−2.0	−1.00	0.0	0.00	8.0
BILL_AMT1	30000.0	51223.330900	73635.860576	−165580.0	3558.75	22381.5	67091.00	964511.0
BILL_AMT2	30000.0	49179.075167	71173.768783	−69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	30000.0	47013.154800	69349.387427	−157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	30000.0	43262.948967	64332.856134	−170000.0	2326.75	19052.0	54506.00	891586.0
BILL_AMT5	30000.0	40311.400967	60797.155770	−81334.0	1763.00	18104.5	50190.50	927171.0
BILL_AMT6	30000.0	38871.760400	59554.107537	−339603.0	1256.00	17071.0	49198.25	961664.0
PAY_AMT1	30000.0	5663.580500	16563.280354	0.0	1000.00	2100.0	5006.00	873552.0
PAY_AMT2	30000.0	5921.163500	23040.870402	0.0	833.00	2009.0	5000.00	1684259.0
PAY_AMT3	30000.0	5225.681500	17606.961470	0.0	390.00	1800.0	4505.00	896040.0
PAY_AMT4	30000.0	4826.076867	15666.159744	0.0	296.00	1500.0	4013.25	621000.0
PAY_AMT5	30000.0	4799.387633	15278.305679	0.0	252.50	1500.0	4031.50	426529.0
PAY_AMT6	30000.0	5215.502567	17777.465775	0.0	117.75	1500.0	4000.00	528666.0
default payment next month	30000.0	0.221200	0.415062	0.0	0.00	0.0	0.00	1.0

whether the person defaulted on the next month's payment (1 = Yes, 0 = No). Some brief descriptive info on the variables is presented in the table below using the pandas describe function.

Column 1 is a unique person ID, column 2 is the person's balance limit, column 3,4,5 and 6 are the person's sex, education level, marriage status, and age respectively. Columns 7-12 contain a history of past payments on a monthly level for the past 6 months coded based on what their payment status is like. Columns 13-18 have their monthly bill totals for the past 6 months. Columns 19-24 contain the dollar amounts they have repaid off of their bill on a monthly level for the past 6 months. After a brief EDA our plan is to test several different types of models to help predict whether or not the person is going to default on their bill. We hope to test a Support Vector Machine, a Random Forest and a Naïve Bayes Model. After looking at the misclassification rates for the 3 basic models we start we will choose 1 type of model to then optimize for performance and try to minimize the misclassification rate.

Judgement

We are going to use the confusion matrix method to test the misclassification rates in our approach. We may also implement other approaches to enhance the accuracy of our result, such as: ROC, AUC, Gini coefficient and K-S test.

Reference Materials

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480

Schedule

08/04-08/07 reading reference materials; learning and exercising network and algorithms

08/08-08/10 coding

08/11-08/13 report & presentation